Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge

Hugo J. Kuijf, J. Matthijs Biesbroek, Jeroen de Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M. Jorge Cardoso, Adrià Casamitjana, D. Louis Collins, Mahsa Dadar, Achilleas Georgiou, Mohsen Ghafoorian, Dakai Jin, April Khademi, Jesse Knight, Hongwei Li, Xavier Lladó, Miguel Luna, Qaiser Mahmood, Richard McKinley, Alireza Mehrtash, Sébastien Ourselin, Bo-Yong Park, Hyunjin Park, Sang Hyun Park, Simon Pezold, Elodie Puybareau, Leticia Rittner, Carole H. Sudre, Sergi Valverde, Verónica Vilaplana, Roland Wiest, Yongchao Xu, Ziyue Xu, Guodong Zeng, *Student Member, IEEE*, Jianguo Zhang, Guoyan Zheng, *Member, IEEE*, Christopher Chen, Wiesje van der Flier, Frederik Barkhof, Max A. Viergever, *Fellow, IEEE*, and Geert Jan Biessels

Manuscript received January 24, 2019; revised March 11, 2019; accepted March 13, 2019. Date of publication March 19, 2019; date of current version October 25, 2019. The work of H. J. Kuijf was supported by The Netherlands Organization for Health Research and Development (ZonMW) through the Off Road Grant under Grant 451001007. The work of S. Andermatt was supported by the MIAC AG, Basel, Switzerland. The work of M. Bento and L. Rittner was supported in part by the Hotchkiss Brain Institute and in part by CAPES process PVE under Grant 88881.062158/2014-01. The work of A. Casamitjana was supported in part by the Spanish Ministerio de Economía y Competitividad through the project MALEGRA under Grant TEC2016-75976-R, in part by the European Regional Development Fund (ERDF), and in part by the Spanish Ministerio de Educacin, Cultura y Deporte FPU Research Fellowship. The work of D. Jin and Z. Xu was supported by the National Institute of Allergy and Infectious Diseases, USA, through the Intramural Research Program. The work of A. Khademi and J. Knight was supported in part by the Natural Science and Engineering Research Council of Canada (NSERC CGS-M) and in part by the Ontario Ministry of Advanced Education and Skills Development (OGS-M). The work of H. Li and J. Zhang was supported by the National Natural Science Foundation of China under Grant 61628212. The work of X. Lladó and S. Valverde was supported in part by the Ministerio de Ciencia y Tecnología, Spain, under Grant TIN2014-55710-R and Grant DPI2017-86696-R. The work of M. Luna and S. H. Park was supported by the National Research Foundation of Korea (NRF) through the Basic Science Research Program funded by the Ministry of Education under Grant 2018R1D1A1B07044473. The work of R. McKinley and R. Wiest was supported by the Swiss Multiple Sclerosis Society. The work of A. Mehrtash was supported in part by the U.S. National Institutes of Health under Grant P41EB015898, in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and in part by the Canadian Institutes of Health Research (CIHR). The work of C. H. Sudre was supported by the Alzheimer's Society Junior Research Fellowship under Grant AS-JF-17-011. The work of V. Vilaplana was supported in part by the Spanish Ministerio de Economía y Competitividad through the project MALEGRA under Grant TEC2016-75976-R and in part by the European Regional Development Fund (ERDF). The work of G. Zeng and G. Zheng was supported in part by the Swiss National Science Foundation under Project 205321_163224. The work of F. Barkhof was supported by the NIHR UCLH Biomedical Research Centre. The work of G. J. Biessels was supported by The Netherlands Organization for Scientific Research (NWO) through the VICI under Grant 918.16.616. *(Corresponding author: Hugo Kuijf.)*

Please see the Acknowledgment section of this paper for the author affiliations.

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

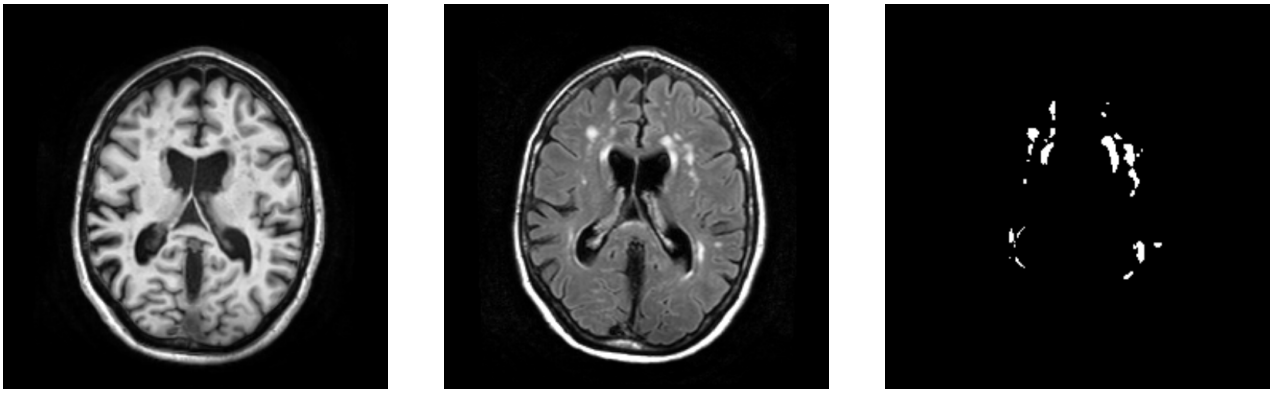
Digital Object Identifier 10.1109/TMI.2019.2905770

***Abstract*— Quantification of cerebral white matter hyperintensities (WMH) of presumed vascular origin is of key importance in many neurological research studies. Currently, measurements are often still obtained from manual segmentations on brain MR images, which is a laborious procedure. The automatic WMH segmentation methods exist, but a standardized comparison of the performance of such methods is lacking. We organized a scientific challenge, in which developers could evaluate their methods on a standardized multi-center/-scanner image dataset, giving an objective comparison: the WMH Segmentation Challenge. Sixty T1 + FLAIR images from three MR scanners were released with the manual WMH segmentations for training. A test set of 110 images from five MR scanners was used for evaluation. The segmentation methods had to be containerized and submitted to the challenge organizers. Five evaluation metrics were used to rank the methods: 1) Dice similarity coefficient; 2) modified Hausdorff distance (95th percentile); 3) absolute log-transformed volume difference; 4) sensitivity for detecting individual lesions; and 5) F1-score for individual lesions. In addition, the methods were ranked on their inter-scanner robustness; 20 participants submitted their methods for evaluation. This paper provides a detailed analysis of the results. In brief, there is a cluster of four methods that rank significantly better than the other methods, with one clear winner. The inter-scanner robustness ranking shows that not all the methods generalize to unseen scanners. The challenge remains open for future submissions and provides a public platform for method evaluation.**

***Index Terms*— Magnetic resonance imaging (MRI), brain, evaluation and performance, segmentation.**

I. Introduction

WHITE matter hyperintensities (WMH) of presumed vascular origin are one of the main manifestations of cerebral small vessel disease and play a key role in stroke,



(a) T1-weighted image (b) FLAIR image (c) Manual WMH segmentation

Fig. 1. Example brain MR images of a subject with white matter hyperintensities (WMH) of presumed vascular origin. On the T1-weighted image (a), WMH show as hypointense regions within the white matter. On the FLAIR image (b), WMH are clearly visible as hyperintense regions within the white matter. The corresponding manual WMH segmentation is shown in (c).

dementia, and ageing [1], [2]. On T2-weighted and fluid-attenuated inversion recovery (FLAIR) brain MR images, WMH are clearly visible as hyperintense regions within the white matter [3]. An example image is shown in Fig. 1, with the manual segmentation shown in Fig. 1(c).

Quantification of WMH is of importance in clinical research studies, where measures of WMH volume, shape, and location are obtained from detailed segmentations. These measures are associated with the presence and severity of clinical symptoms, such as cognitive impairment and gait disturbances, and are likely to find their way into daily clinical practice, supporting diagnosis, prognosis, and treatment monitoring [2], [4]. However, manual delineation of WMH is a time-consuming and observer-dependent procedure.

Automatic WMH segmentation methods have been developed, but a review by Caligiuri *et al.* [5] revealed a key issue: it is hard to compare the various methods that are described in the literature. Each proposed segmentation method has been evaluated on a different ground truth (different number of subjects, different experts, different protocols), using different evaluation criteria.

A further challenge of automatic WMH segmentation methods is the deployment of such a method within a new institute that might have different scanners or imaging protocols. Many (deep) machine learning methods require some form of transfer learning or fine-tuning on the target images [6], which in practice is not always feasible.

These issues are not unique to the task of automatic WMH segmentation, but occur in many medical image analysis tasks. Organizing a scientific challenge is a way to address this, having a number of competing methods perform the same task on the same data. This has been successfully applied to various tasks, such as liver segmentation [7], image registration [8], coronary calcium scoring [9], or gland segmentation in histology images [10]. In the past, a number of challenges have been organized that included abnormalities on brain MR images, such as the multiple sclerosis (MS) lesions [11], [12], tumor [13], or tissue [14] segmentation challenges.[1] However, none of these challenges focuses on WMH of presumed vascular origin (although MS lesions share some characteristics with such WMH; and the brain tissue segmentation challenge included WMH lesions, but not as a separate task).

The WMH Segmentation Challenge described in this paper provides a standardized assessment of automatic methods for the segmentation of WMH. The task for the challenge was defined as: "the segmentation of white matter hyperintensities of presumed vascular origin on brain MR images"[2] [3]. Key features of the challenge include: Participants have to submit their method to the organizers for independent evaluation on a test set. The test set includes data from two additional scanners not in the training data, to evaluate generalizability of segmentation methods across scanners. The dataset was derived from patients with various degrees of ageing related degenerative and vascular pathologies, which is important for the generalizability since segmentation methods should be able to deal with this variation. Evaluation is performed using five different metrics and participants are ranked relative to each other.

In this paper, the organization of the challenge, its results, and a detailed evaluation are presented.

II. Methods

A. Training and Test Data

A total of 60 training and 110 test images were used in this challenge. Imaging data was acquired from five different scanners, from three different vendors, in three different institutes: the University Medical Center (UMC) Utrecht, VU University Medical Centre (VU) Amsterdam, both in the Netherlands, and the National University Health System (NUHS) in Singapore. For each subject, a 3D T1-weighted and a 2D multi-slice FLAIR image were provided.

The training data consisted of sixty images: twenty 3 T images of a single scanner of each institute. The test set included ninety images (three times thirty) of those same scanners and additionally twenty images (two times ten) of

[1]For a more complete overview, visit: https://grand-challenge.org/challenges/

[2]https://wmh.isi.uu.nl/details/

TABLE I
OVERVIEW OF THE NUMBER OF IMAGES AVAILABLE FOR TRAINING (TR.) AND TEST (TE.)

Institute	Scanner	Tr.	Te.
UMC Utrecht	3 T Philips Achieva	20	30
NUHS Singapore	3 T Siemens TrioTim	20	30
VU Amsterdam	3 T GE Signa HDxt	20	30
	1.5 T GE Signa HDxt	0	10
	3 T Philips Ingenuity (PET/MR)	0	10

scanners that were not in the training data set. An overview of the data set is given in Table I.

Subjects included from UMC Utrecht and VU Amsterdam were selected from the memory clinic patients of both institutes [15].

Subjects included from the NUHS Singapore were selected from the Memory Ageing and Cognition Centre Cohort recruited from the memory clinics of the National University Hospital and St. Luke's Hospital in Singapore [16].

For each scanner, subjects were randomly picked from all subjects and randomly placed into the training or test sets.

1) MRI Parameters: All 3D sequences were acquired in the sagittal direction and all 2D multi-slice sequences in the transversal direction.

UMC Utrecht, 3 T Philips Achieva: 3D T1-weighted sequence (192 slices, voxel size: $1.00 \times 1.00 \times 1.00$ mm^3, repetition time (TR)/echo time (TE): 7.9/4.5 ms), 2D FLAIR sequence (48 slices, voxel size: $0.96 \times 0.95 \times 3.00$ mm^3, TR/TE/inversion time (TI): $11,000/125/2,800$ ms)

NUHS Singapore, 3 T Siemens TrioTim: 3D T1-weighted sequence (voxel size: $1.00 \times 1.00 \times 1.00$ mm^3, TR/TE/TI: $2,300/1.9/900$ ms), 2D FLAIR sequence (voxel size: $1.00 \times 1.00 \times 3.00$ mm^3, TR/TE/TI: $9,000/82/2,500$ ms)

VU Amsterdam, 3 T GE Signa HDxt: 3D T1-weighted sequence (176 slices, voxel size: $0.94 \times 0.94 \times 1.00$ mm^3, TR/TE: 7.8/3.0 ms), 3D FLAIR sequence (132 slices, voxel size: $0.98 \times 0.98 \times 1.20$ mm^3, TR/TE/TI: $8,000/126/2,340$ ms)

VU Amsterdam, 1.5 T GE Signa HDxt: 3D T1-weighted sequence (172 slices, voxel size: $0.98 \times 0.98 \times 1.50$ mm^3, TR/TE: 12.3/5.2 ms), 3D FLAIR sequence (128 slices, voxel size: $1.21 \times 1.21 \times 1.30$ mm^3, TR/TE/TI: $6,500/117/1,987$ ms)

VU Amsterdam, 3 T Philips Ingenuity (PET/MR): 3D T1-weighted sequence (180 slices, voxel size: $0.87 \times 0.87 \times 1.00$ mm^3, TR/TE: 9.9/4.6 ms), 3D FLAIR sequence (321 slices, voxel size: $1.04 \times 1.04 \times 0.56$ mm^3, TR/TE/TI: $4,800/279/1,650$ ms)

All 3D FLAIR sequences were resampled into the transversal direction with slices of 3 mm thickness for two reasons: (1) to save time on the manual annotation of WMH and (2) to become more similar to the 2D multi-slice sequences.

An example FLAIR image of each scanner is shown in Appendix A Figure 7.[3]

2) Data Pre-Processing: All images were bias-corrected using SPM12 [17]. Using the `elastix` toolbox for image registration [18], the 3D T1-weighted images were aligned with the (resampled) FLAIR images. The transformation parameters were provided with the data. The faces of the subjects were manually removed from all sequences and the masks used for that were provided as well.

Data before and after preprocessing is provided on the challenge website for registered participants: https://wmh.isi.uu.nl/data/.

3) Manual Reference Standard: WMH and other pathologies (i.e. lacunes and non-lacunar infarcts, (micro)hemorrhages) were manually segmented in accordance with the STandards for ReportIng Vascular changes on nEuroimaging (STRIVE) criteria [3]. The outline of WMH and other pathology was delineated using a contour drawing technique by an expert observer (O1). This observer had extensive prior experience with the manual segmentation of WMH and had segmented 1000+ cases before this dataset. Manual delineations were peer-reviewed by a second expert observer (O2) with eleven years of experience in quantitative neuroimaging and clinical neuroradiology. In case of mistakes, errors, or delineations that were not according to the STRIVE criteria, O1 corrected the manual segmentation in a consensus meeting with O2. Hence, the provided reference standard is the corrected segmentation of O1, after peer review by O2.

The contours were converted to binary masks, whereby all voxels whose volume was within the manual delineation for >50 %, were considered WMH. Background received label 0 and WMH label 1. Other pathology was converted to binary masks as well, receiving label 2. These masks were dilated by 1 pixel in-plane (with a $3 \times 3 \times 1$ voxel kernel). In case of overlap between labels 1 and 2 (after dilation), label 1 was assigned.

Two additional observers segmented the sixty training images to obtain inter-observer agreement measures. Observer O3 was trained for WMH segmentation, but had no extensive prior experience. Observer O4 was trained for WMH segmentation and had prior experience.

B. Set-Up of the Challenge

Participants could register on the challenge website and download the training data. Methods had to be containerized with Docker[4] [19] and submitted for evaluation. Containerization eases deployment of methods and guarantees that the method will produce identical output when run on a different platform. To ensure this, the output of the containerized method on the first training subject was sent back to the participants for verification.

During testing, the containerized method was run on each test subject one by one. No identifiers were present that would indicate from which of the five scanners the current subject originated. After processing a subject, the container was completely destroyed and reloaded. Full details on how

[3] Available in the supplementary files / multimedia tab.

[4] https://www.docker.com/

the containers would be run, including a Python and MATLAB example container, were provided on the challenge website.[5]

An NVIDIA Titan Xp GPU was available for methods that needed one.

C. Participants

Twenty teams submitted their method before the deadline and participated in the challenge. A brief summary of each method is given below, in alphabetical order.

achilles a neural network similar to HighResNet [20] and DeepLab v3 [21], utilizing atrous (dilated) convolutions, atrous spatial pyramid pooling, and residual connections. The network is trained only on the FLAIR images, taking random 71^3 sized patches, and applying scaling and rotation augmentations [22].

cian a network based on multi-dimensional gated recurrent units (MD-GRU) was trained on 3D patches using data augmentation techniques including random deformation, rotation and scaling [23]–[25].

hadi a random forest classifier trained on multi-modal image features. These include intensities, gradient, and Hessian features of the original images, after smoothing, and of generated super-voxels [26].

ipmi-bern a two-stage approach that uses fully convolutional neural networks to first extract the brain from the images and second identifies WMH within the brain. Both stages implement long and short skip connections. The second stage produces output at three different scales. Data augmentation was applied, including rotations and mirroring [27].

k2 a 2D fully convolutional neural network with an architecture similar to U-Net [28]. A number of models were trained for the whole dataset, as well as for each individual scanner. During application, first the type of scanner was predicted and next that specific model was applied together with the model trained on all data [29].

knight a voxel-wise logistic regression model that is fitted independently for each voxel in the FLAIR image. Images were transformed to the MNI-152 standard space [30] for training and at test time the parameter maps were warped to the subject space [31], [32].

lrde a modification of the pre-trained 16-layer VGG network [33], where the FLAIR, T1, and a high-pass filtered FLAIR are used as multi-channel input. The VGG network had its fully connected layers replaced by a number of convolutional layers [34]–[37].

misp a 3D convolutional neural network with 18 layers using patches of $27 \times 27 \times 9$ voxels. The first eight layers were trained separately for the FLAIR and T1 images and had skip-connections [38] [39].

neuro.ml a neural network using the DeepMedic [40] architecture, having two parallel branches that process the images at two different scales. The network used 3D patches, which were sampled such that 60 % of the patches contained a WMH [41].

nic-vicorob a 10-layer 3D convolutional neural network architecture previously used to segment multiple sclerosis lesions [42]. A cascaded training procedure was employed, training two separate networks to first identify candidate lesion voxels and next to reduce false positive detections. A third network re-trains the last fully connected layer to perform WMH segmentation [43].

nih_cidi a fully convolutional neural network modified from the U-Net architecture [28] was used to segment WMH on the FLAIR images. Next, another network was trained to segment the white matter from T1 images, and the segmented white matter mask is applied to remove false positives from the WMH segmentation results. The original U-Net architecture was trimmed to keep only three pooling layers [44].

nist a random decision forest classifier trained on location and intensity features [45]–[47].

nlp_logix a multiscale deep neural network similar to [48], with some minor modifications and no spatial features. The network was trained in ten folds and the three best performing checkpoints on the training data were selected. These were applied on the test set and the results averaged [49].

scan a densely connected convolutional network using dilated convolutions [50], [51]. In each dense block, the output is concatenated to the input before passing it to the next layer. Two classifiers were trained: one to apply brain extraction and the second to find lesions within the extracted brain [52].

skkumedneuro an intensity-based thresholding method with region growing approach to segment periventricular and deep WMH separately, and two random forest classifiers for false positive reduction. Per imaging modality, 19 texture and 100 "multi-layer" features were computed. The "multi-layer" features were computed using a feed-forward convolutional network with fixed filters (e.g. averaging, Gaussian, Laplacian); consisting of two convolutional, two max-pooling, and one fully connected layer [53].

sysu_media a fully convolutional neural network similar to U-Net [28]. An ensemble of three networks was trained with different initializations. Data normalization and augmentation was applied. To remove false positive detections, WMH in the first and last $\frac{1}{8}$th slices was removed [54], [55].

text_class a random forest classifier trained primarily on texture features. Features include local binary pattern, structural and morphological gradients, and image intensities [56], [57].

tig a three-level Gaussian mixture model, slightly adapted from [58]. The model is iteratively modified and evaluated, until it converges. After that, candidate WMH is selected and possible false positives are pruned based on their location [59].

tignet a neural network with the HighResNet architecture [20]. The network was trained on $2,660$ images segmented using the previous method of team **tig** [58], [60].

upc_dlmi a neural network modified from the V-Net architecture [61]. An additional network with convolutional layers is trained on upsampled images and then concatenated with the output of the V-Net [62].

Detailed information on each method can be found online at https://wmh.isi.uu.nl/results/results-miccai-2017/.

[5] https://wmh.isi.uu.nl/methods/

D. Evaluation and Ranking

Methods were evaluated according to five criteria: (1) the Dice Similarity Coefficient (DSC), (2) a modified Hausdorff distance (95th percentile; H95), (3) the absolute percentage volume difference (AVD), (4) the sensitivity for detecting individual lesions (recall), and (5) F1-score for individual lesions (F1). For recall and F1, individual lesions are defined as 3D connected components within an image. The exact implementation of each metric was put online[6] beforehand and could be used by participants for self-evaluation during development. During evaluation of the results, it was discovered that the AVD metric had a slight flaw. A method could undersegment WMH by at most 100%, but could oversegment WMH almost infinitely. Therefore, in this manuscript, the AVD metric was replaced by the absolute log-transformed volume difference (lAVD, (1)).

$$\text{lAVD} = |\log \frac{\text{segmented volume}}{\text{true volume}}| \quad (1)$$

The final ranking was based on the five metrics and each method received a rank relative to the performance of all methods. This was computed in a number of steps. First, each metric was averaged over all test scans per method. For each metric, the methods were sorted from best to worst. Next, the best method received a rank of 0 and the worst method a rank of 1; all other methods were ranked relatively in the range $(0, 1)$. Finally, the five ranks were averaged into the overall rank.

95% confidence intervals on each individual metric and the final ranking were computed using bootstrapping. The bootstrap distribution included $2,000$ samples taken randomly from the test set with replacement. Non-overlapping confidence intervals indicate a significant difference between methods, with $\alpha = 0.05$.

It is expected that methods might have more difficulties detecting and segmenting small lesions compared to large lesions. For each subject, the recall will be computed separately for individual lesions smaller than or equal to the median lesion size and for lesions larger than the median lesion size.

Additionally, a ranking was computed based solely on the inter-scanner differences. This ranking highlights which methods have the most robust performance across various scanners. For each method and scanner, the median performance of each metric was computed. Next, the standard deviation of those medians per scanner was averaged; giving a single value per metric: the standard deviation of the median per scanner. Methods were then ranked based on this value: first per metric and then averaged over all five metrics. A lower standard deviation across the median performance on all scanners (for all metrics) indicates a better inter-scanner robustness.

Finally, the Simultaneous Truth And Performance Level Estimation (STAPLE) algorithm [63] was applied to all methods and to the top-ranking methods. STAPLE takes multiple segmentations as input and produces a combined segmentation, which was evaluated and ranked separately. It has been shown for other applications, e.g. brain tumor segmentation [13], that fusing the output of multiple methods can outperform all individual methods.

[6]https://github.com/hjkuijf/wmhchallenge/blob/master/evaluation.py

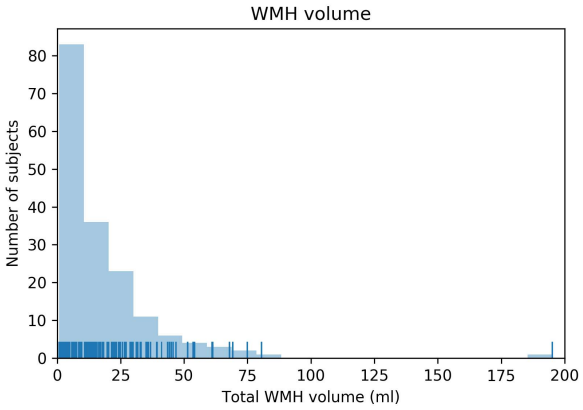


Fig. 2. Histogram showing the WMH volume distribution throughout the dataset. The ticks on the x-axis represent each individual subject.

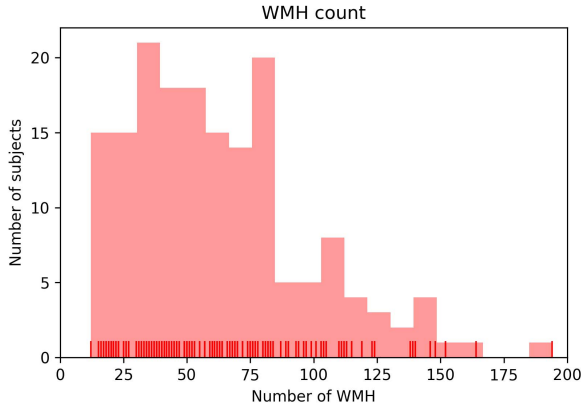


Fig. 3. Histogram showing the WMH count distribution throughout the dataset. The ticks on the x-axis represent each individual subject. An individual lesion is defined as a 3D connected component within an image.

III. Results

The subjects included in the challenge were (mean ± sd) 70.1 ± 9.3 years old and 50 % were male. The WMH volume in the dataset was (mean ± sd): 16.9 ± 21.6 ml (min: 0.78 ml, Q1: 3.24 ml, median: 11.18 ml, Q3: 23.00 ml, max: 195.15 ml; see Fig. 2). The WMH count in the dataset was (mean ± sd): 62 ± 35 lesions (min: 12 lesions, Q1: 36 lesions, median: 57 lesions, Q3: 81 lesions, max: 194 lesions; see Fig. 3). The distribution of lesions throughout the dataset is shown in the top row of Fig. 4. There were no significant differences between the training and test sets for age ($p = 0.45$), gender ($p = 0.87$), WMH volume ($p = 0.74$), WMH count ($p = 0.75$), the presence of lacunes ($p = 0.86$), nor for the volume of other pathology ($p = 0.62$). Tests for age and volumes were performed using Welch's unequal variances t-test [64]. Tests for gender and presence of lacunes were performed using Fisher's exact test.

The bottom row of Table II shows the inter-observer agreement of observers O3 and O4 compared with the manual reference standard of the sixty training images. Additionally,

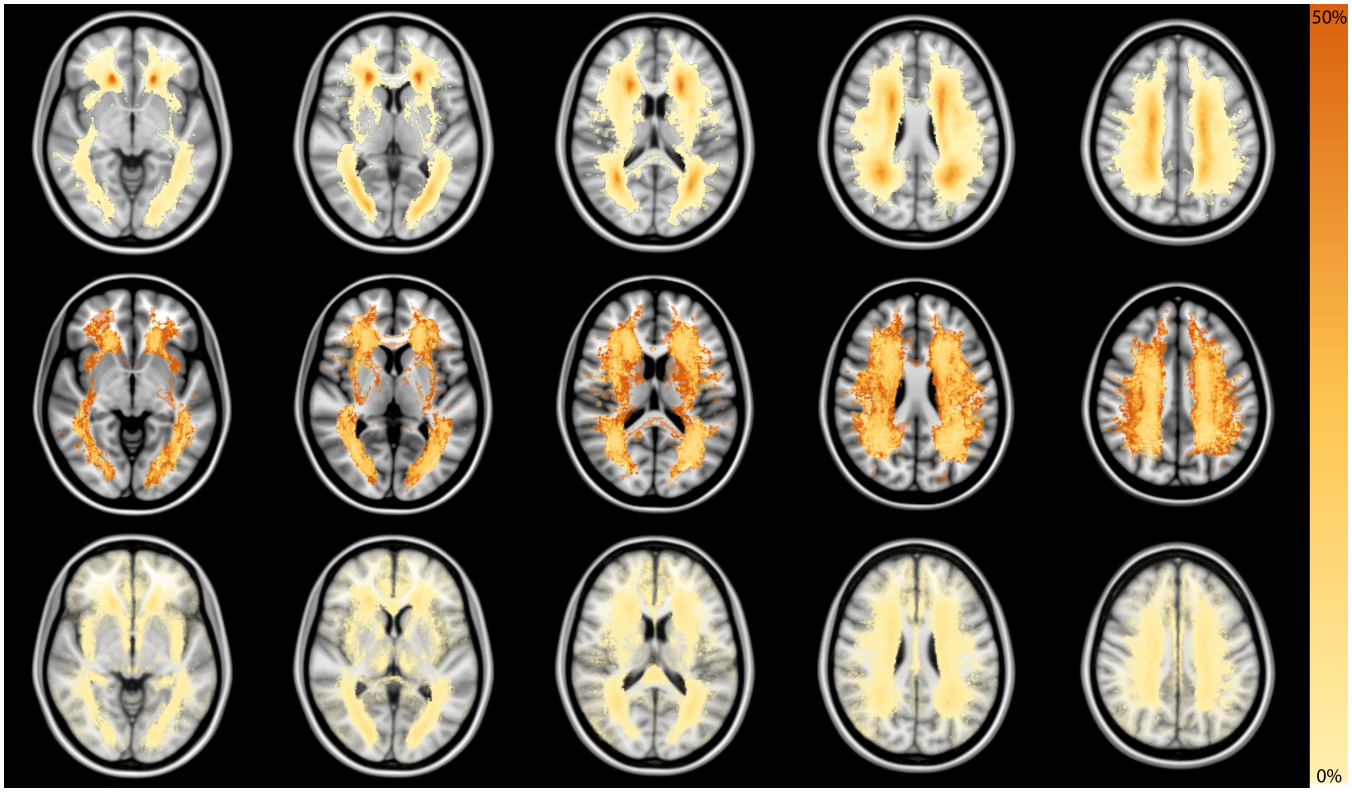


Fig. 4. The MNI-152 standard brain template [30], showing different overlays. Top row: WMH distribution throughout the dataset, where the color indicates the percentage of subjects that have a lesion in that specific voxel. Middle row: false negative rate, showing the percentage of lesions that were missed in a specific voxel. Bottom row: false positive rate, showing the percentage of false positives in a specific voxel. All voxels where only one subject has a lesion are shown half translucent.

the associated positions of O3 and O4 with respect to all methods in the ranking is provided. This is the position these observers would have achieved if they had participated as method in the challenge.

The mean performance of each participating method on each individual metric is shown in Table II, together with the 95% confidence intervals. The method of **sysu_media** performed best on the DSC, H95, and recall metrics. The method of **cian** performed best on the lAVD metric. The method of **nlp_logix** performed best on the F1 metric. Fig. 5 shows boxplots of all results of each method on each metric.

The final ranking is shown in Table III, together with the 95% confidence intervals.

The middle and bottom rows of Fig. 4 show spatial maps of the false negative rate and false positive rate, respectively, of all methods combined. Appendix C Figures 8–27[7] show these spatial maps per method, ordered by their final ranking.

Fig. 6 shows the (relative) difference in recall between small and large lesions. All methods perform worse in recalling small lesions compared to large lesions. For example, the method of **sysu_media** has a recall for large lesions of 94% and a recall for small lesions of 76%, resulting in a relative difference of −20%. Overall, the drop in recall ranges from −20% (**sysu_media**) to −87% (**text_class**), as indicated by the solid lines in the figure.

[7]Available in the supplementary files / multimedia tab.

Table IV highlights various properties of all methods, sorted by their final ranking. The top 11 methods all employ some form of deep learning, with a U-Net-like architecture [28] being the overall most common. Amongst the non-deep learning methods, the use of a random forest classifier is most common. Almost all methods apply various kinds of pre-processing, where normalizing intensities of an image to a standardized range is applied by most methods. Some methods apply post-processing techniques, mainly aimed at reducing the number of false positive detections. The H95 and F1 metrics are most sensitive to false positive detections, but the methods that apply post-processing do not have a notable better score on these metrics than nearby ranking methods. When considering only deep learning methods, most use data augmentation to generate more training samples. Scaling, rotating, and mirroring an image are quite common, but the top 2 methods also apply shearing or non-linear deformations. The last columns of Table IV highlight some properties of deep learning methods, in which a few clusters can be distinguished. Top ranking methods have applied dropout during training, some form of hard negative mining, and use an ensemble of networks. Three methods use dilated convolutions, but these cluster in the middle of the ranking. Most methods that use 3D convolutions appear to rank at the bottom. Using batch normalization, multi scale approaches, or learning rate schedules does not seem to influence the ranking; and neither does the choice of loss function.

TABLE II

MEAN PERFORMANCE AND 95 % CONFIDENCE INTERVALS OF EACH PARTICIPATING METHOD ON EACH INDIVIDUAL METRIC. METRICS INCLUDE: (1) DICE SIMILARITY COEFFICIENT (DSC), (2) MODIFIED HAUSDORFF DISTANCE (95TH PERCENTILE; H95), (3) ABSOLUTE OF THE PERCENTAGE VOLUME DIFFERENCE (AVD), (4) SENSITIVITY FOR DETECTING INDIVIDUAL LESIONS (RECALL), AND (5) F1-SCORE FOR INDIVIDUAL LESIONS (F1). BOLD INDICATES THAT A METHOD HAS THE BEST SCORE ON THAT METRIC. METHODS ARE SORTED BASED ON THE FINAL RANKING AS SHOWN IN TABLE III. THE BOTTOM ROWS INCLUDE THE RESULTS OF THE SIMULTANEOUS TRUTH AND PERFORMANCE LEVEL ESTIMATION (STAPLE) ALGORITHM APPLIED ON ALL METHODS AND ON THE TOP 4 RANKING METHODS, AND THE RESULTS OF OBSERVERS O3 AND O4, TOGETHER WITH THE ASSOCIATED POSITIONS IN THE RANKING IF STAPLE, O3, AND O4 WOULD HAVE PARTICIPATED IN THE CHALLENGE. NOTE THAT O3 AND O4 SEGMENTED THE SIXTY TRAINING IMAGES

#	Team	DSC	H95 (mm)	lAVD	Recall	F1
1	sysu_media	**0.80 (0.78 - 0.82)**	**6.30 (4.75 - 7.93)**	0.193 (0.165 - 0.224)	**0.84 (0.82 - 0.86)***	0.76 (0.73 - 0.78)†
2	cian	0.78 (0.76 - 0.80)	6.82 (4.92 - 9.22)	**0.193 (0.162 - 0.228)**	0.83 (0.81 - 0.84)*	0.70 (0.67 - 0.73)‡
3	nlp_logix	0.77 (0.75 - 0.80)	7.16 (5.61 - 8.82)	0.219 (0.174 - 0.271)	0.73 (0.71 - 0.76)**	**0.78 (0.76 - 0.80)**†
4	nic-vicorob	0.77 (0.74 - 0.79)	8.28 (6.60 - 10.06)	0.248 (0.201 - 0.303)	0.75 (0.73 - 0.77)**	0.71 (0.68 - 0.73)‡
5	k2	0.77 (0.74 - 0.79)	9.79 (7.72 - 12.28)	0.246 (0.187 - 0.310)	0.59 (0.56 - 0.61)	0.70 (0.68 - 0.72)‡
6	misp	0.72 (0.69 - 0.75)	14.88 (10.52 - 19.41)	0.258 (0.167 - 0.388)	0.63 (0.60 - 0.65)	0.68 (0.65 - 0.70)‡
7	lrde	0.73 (0.70 - 0.76)	14.54 (10.32 - 19.31)	0.309 (0.218 - 0.442)	0.63 (0.60 - 0.66)	0.67 (0.65 - 0.69)‡
8	nih_cidi	0.68 (0.65 - 0.70)	12.82 (10.54 - 15.16)	0.281 (0.200 - 0.394)	0.59 (0.56 - 0.62)	0.54 (0.51 - 0.57)
9	ipmi-bern	0.69 (0.67 - 0.72)	9.72 (7.98 - 11.56)	0.225 (0.178 - 0.275)	0.44 (0.42 - 0.46)	0.57 (0.55 - 0.58)
10	scan	0.63 (0.59 - 0.66)	14.34 (12.25 - 16.50)	0.277 (0.223 - 0.336)	0.55 (0.52 - 0.58)	0.51 (0.48 - 0.53)
11	achilles	0.63 (0.60 - 0.66)	11.82 (9.80 - 13.94)	0.276 (0.226 - 0.331)	0.45 (0.42 - 0.47)	0.52 (0.50 - 0.53)
12	skkumedneuro	0.58 (0.54 - 0.61)	19.02 (16.64 - 21.58)	0.384 (0.292 - 0.503)	0.47 (0.44 - 0.49)	0.51 (0.48 - 0.54)
13	tignet	0.59 (0.56 - 0.63)	21.58 (18.15 - 25.33)	0.533 (0.450 - 0.623)	0.46 (0.41 - 0.51)	0.45 (0.42 - 0.49)
14	tig	0.60 (0.56 - 0.63)	17.86 (15.57 - 20.20)	0.400 (0.333 - 0.474)	0.38 (0.36 - 0.41)	0.42 (0.40 - 0.44)
15	knight	0.70 (0.67 - 0.72)	17.03 (14.48 - 19.88)	0.352 (0.290 - 0.427)	0.25 (0.22 - 0.27)	0.35 (0.32 - 0.38)
16	upc_dlmi	0.53 (0.48 - 0.58)	27.01 (22.25 - 31.99)	0.612 (0.481 - 0.762)	0.57 (0.53 - 0.60)	0.42 (0.38 - 0.46)
17	nist	0.53 (0.49 - 0.57)	15.91 (14.44 - 17.42)	0.581 (0.469 - 0.695)	0.37 (0.34 - 0.40)	0.25 (0.22 - 0.27)
18	neuro.ml	0.51 (0.45 - 0.56)	37.36 (33.70 - 40.89)	1.033 (0.836 - 1.241)	0.71 (0.68 - 0.75)**	0.21 (0.19 - 0.24)
19	text_class	0.50 (0.45 - 0.54)	28.23 (24.15 - 32.68)	0.605 (0.492 - 0.724)	0.27 (0.25 - 0.29)	0.29 (0.26 - 0.31)
20	hadi	0.23 (0.19 - 0.27)	52.02 (49.25 - 54.82)	1.685 (1.448 - 1.939)	0.58 (0.52 - 0.63)	0.11 (0.09 - 0.12)
4	STAPLE (all)	0.77 (0.74 - 0.80)	**5.74 (4.26 - 7.43)**	0.315 (0.249 - 0.393)	0.77 (0.75 - 0.79)	0.74 (0.71 - 0.76)
2	STAPLE (top 4)	**0.80 (0.78 - 0.82)**	6.43 (4.48 - 8.81)	**0.171 (0.144 - 0.201)**	0.80 (0.78 - 0.82)	0.76 (0.74 - 0.78)
5	O3	0.77 (0.74 - 0.80)	6.79 (5.32 - 8.54)	**0.176 (0.135 - 0.222)**	0.65 (0.62 - 0.69)	0.74 (0.71 - 0.76)
4	O4	0.79 (0.76 - 0.81)	7.22 (5.36 - 9.36)	0.195 (0.148 - 0.245)	0.66 (0.63 - 0.70)	0.76 (0.73 - 0.78)

* **sysu_media** and **cian** perform significantly better on the recall metric than all other teams.
** **nic-vicorob**, **nlp_logix**, and **neuro.ml** perform significantly better on the recall metric than all remaining teams.
† **nlp_logix** and **sysu_media** perform significantly better on the F1 metric than all other teams.
‡ **nic-vicorob**, **k2**, **cian**, **misp**, and **lrde** perform significantly better on the F1 metric than all remaining teams.

The inter-scanner robustness was determined as follows: Appendix C Figures 8–27 show the median performance of each method per metric per scanner (the line in the individual boxplots). Per metric, the standard deviation of the median values per scanner is computed. Next, methods are ranked based on those values, where a lower standard deviation indicates better inter-scanner performance. The result of this inter-scanner ranking is shown in the last column of Table III, together with the new position of that method in the ranking. The method of **ipmi-bern** achieves the highest inter-scanner rank and **sysu_media** is just behind on the second rank. The methods of **achilles** and **knight** enter the top 4 of the ranking.

STAPLE was applied on all methods and on the top 4 ranking methods, since these rank significantly higher than all other methods. The results are shown on the bottom rows of Table II and in Appendix C Figures 28 and 29. STAPLE on all methods would rank fourth in the challenge and achieves the best H95. STAPLE on the top 4 ranking methods would rank second in the challenge and achieves the best DSC and lAVD. When re-computing the inter-scanner robustness, both STAPLE methods outperform all other methods. STAPLE compared with the top 3 methods in the inter-scanner ranking is shown separately in Table V, because the relative ranking values change when including STAPLE.

Finally, it could be hypothesized that low ranking methods suffer from training set overfitting [65] or poor generalization. This was evaluated by applying all submitted methods to the training data and comparing the performance on the training data to the performance on the test data. This analysis shows excellent correlation (R-squared: 0.94, with $p < 0.001$), suggesting that there is no indication for overfitting of methods on the training data.

IV. DISCUSSION

We have presented a standardized assessment of automatic methods for the segmentation of white matter hyperintensities of presumed vascular origin. This assessment was performed

TABLE III

FINAL RANKING OF THE METHODS THAT PARTICIPATED IN THE CHALLENGE. THE COLUMN RANK SHOWS THE RELATIVE PERFORMANCE OF EACH METHOD, BASED ON ALL FIVE METRICS LISTED IN TABLE II, TOGETHER WITH THE 95 % CONFIDENCE INTERVALS. THE COLUMN INTER-SCANNER RANK SHOWS THE RANKING WHEN IT IS COMPUTED SOLELY BASED ON INTER-SCANNER ROBUSTNESS. THE SYMBOLS BETWEEN BRACKETS INDICATE WHETHER A TEAM IS RANKED ON THE SAME POSITION (−), LOWER (∨), OR HIGHER (∧) COMPARED WITH THE ORIGINAL RANKING; WITH THE NEW POSITION INDICATED AS WELL. DOTTED LINES INDICATE CLUSTERS OF METHODS THAT RANK SIGNIFICANTLY DIFFERENT FROM METHODS RANKED ABOVE/BELOW, BECAUSE OF NON-OVERLAPPING CONFIDENCE INTERVALS

#	Team	Rank (95 % CI)	Inter-scanner rank
1	sysu_media	0.0068 (0.0019 - 0.0161)†	0.0375 (∨ 2)
2	cian	0.0357 (0.0248 - 0.0539)‡	0.0831 (∨ 5)
3	nlp_logix	0.0520 (0.0365 - 0.0744)‡	0.1111 (∨ 7)
4	nic-vicorob	0.0785 (0.0577 - 0.1045)‡	0.1629 (∨ 11)
5	k2	0.1437 (0.1188 - 0.1711)	0.1174 (∨ 8)
6	misp	0.1740 (0.1356 - 0.2273)	0.1915 (∨ 12)
7	lrde	0.1782 (0.1395 - 0.2290)	0.3510 (∨ 17)
8	nih_cidi	0.2376 (0.2131 - 0.2680)	0.1570 (∨ 10)
9	ipmi-bern	0.2537 (0.2391 - 0.2727)	0.0345 (∧ 1)
10	scan	0.2836 (0.2631 - 0.3099)	0.2252 (∨ 14)
11	achilles	0.3058 (0.2896 - 0.3276)	0.0714 (∧ 3)
12	skkumedneuro	0.3649 (0.3325 - 0.4044)	0.1105 (∧ 6)
13	tignet	0.4090 (0.3765 - 0.4481)	0.2969 (∨ 15)
14	tig	0.4097 (0.3795 - 0.4454)	0.1289 (∧ 9)
15	knight	0.4320 (0.4082 - 0.4598)	0.0785 (∧ 4)
16	upc_dlmi	0.4429 (0.3903 - 0.5016)	0.7415 (∨ 20)
17	nist	0.5040 (0.4724 - 0.5404)	0.3052 (∧ 16)
18	neuro.ml	0.5615 (0.5193 - 0.6084)	0.6110 (∨ 19)
19	text_class	0.5961 (0.5539 - 0.6430)	0.2117 (∧ 13)
20	hadi	0.8886 (0.8687 - 0.9103)	0.4974 (∧ 18)

† sysu_media ranks significantly higher than all other participants.
‡ cian, nlp_logix, and nic-vicorob rank significantly higher than all remaining participants.

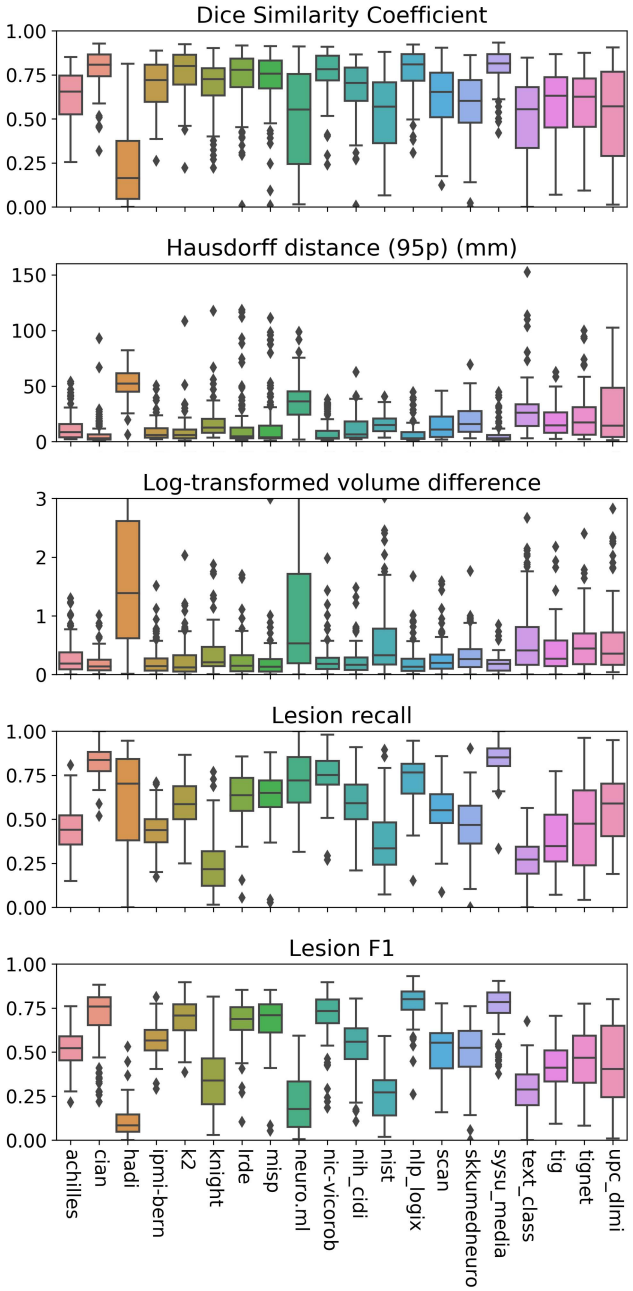


Fig. 5. Boxplots showing all five metrics per method. The box indicates the interquartile range (IQR) with a line at the median. The whiskers extend up to 1.5 times the IQR and the fliers indicate the remaining data points. Note for the Hausdorff distance that **hadi** did not produce any output for 10 subjects and hence their boxplot is based on only 100 subjects (see Appendix C Figure 27 for full details). Note for the log-transformed volume difference that for visibility purposes, this figure is clipped at 3.0. Teams **hadi**, **lrde**, **misp**, **neuro.ml**, **nih_cidi**, **nist**, **skkumedneuro**, **text_class**, and **upc_dlmi** have IAVD values above 3.0. For full details, see Appendix C Figures 27, 14, 13, 25, 15, 24, 19, 26, and 23, respectively.

in the context of the WMH Segmentation Challenge, hosted at the 20th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2017, Québec City, Quebec, Canada.

The manual reference standard was created in consensus by two skilled observers with extensive prior experience in WMH segmentation, which resulted in high quality WMH segmentations. Two additional observers individually segmented the sixty training images, without a consensus reading, to determine inter-observer agreement. The top-ranking methods achieve similar or superior performance as these two individual observers, which suggests that automatic methods might be able to replace individual observers in WMH segmentation. The moderate recall of the individual observers is mainly caused by not segmenting or missing small WMH. The F1 is higher than the recall, which is opposite for most automatic methods, and indicates that both O3 and O4 have hardly segmented any false positive WMH.

The organizers have chosen not to disclose the test set, contrary to what is common in medical image analysis challenges. By keeping the test set secret, a high reliability of the results can be ensured because it obviates the possibility of (visual) self-evaluation by participants.

The rapidly increasing popularity of (deep) neural networks as methodology of choice for analyzing medical images [66] is noticeable in this challenge as well. Fourteen of the twenty

TABLE IV
OVERVIEW OF VARIOUS PROPERTIES OF ALL METHODS. METHODS ARE SORTED BASED ON THE FINAL RANKING AS SHOWN IN TABLE III

									Neural network features[5]							
#	**Team**	**Pre**[1]	**Method**	**Post**[2]	**Data**	**DL**[3]	**Aug**[4]	**Loss function**	**Dim**	**Dil**	**BN**	**Drop**	**MS**	**LR**	**HN**	**Ens**
1	sysu_media	I,R	U-Net	SL	T,F	✓	H,R,S	DSC	2D							✓
2	cian	F,I	MDGRU		T,F	✓	D,R,S	multinom. log.	2D†			✓				
3	nlp_logix	I,S	CNN		T,F	✓		cross-entropy	2D			✓	✓	✓	✓	✓
4	nic-vicorob		CNN	SM	T,F	✓	M,R	cross-entropy	3D		✓	✓			✓	
5	k2	I,R,S	U-Net		T,F	✓	M	DSC	2D			✓		✓		✓
6	misp	I,R	CNN		T,F	✓		mean sq. error	3D		✓	✓			✓	
7	lrde	F,I	VGG-16		T,F	✓	R,S	multinom. log.	2D					✓		
8	nih_cidi	S	U-Net	G	T,F	✓	M,R	cross-entropy	2D		✓			✓		
9	ipmi-bern	I,S	U-Net		T,F	✓	M,R	cross-entropy	2D		✓		✓	✓		
10	scan	S	DenseNet		T,F	✓		cross-entropy	2D	✓						
11	achilles	I,R	HighResNet		F	✓	R,S	DSC	3D	✓	✓		✓			
12	skkumedneuro	I,S	RF		T,F											
13	tignet	B,I,T	HighResNet		T,F*	✓		DSC	3D	✓	✓			✓		
14	tig	B,S,T	GMM	FP	T,F											
15	knight	B,I,S,T	VLR	SM	F*		M,T,Y	DSC								
16	upc_dlmi	I	U-Net		T,F	✓	M	DSC	3D		✓		✓	✓		
17	nist	B,I,T	RF		T,F											
18	neuro.ml		DeepMedic		T,F	✓		cross-entropy	3D				✓	✓		
19	text_class	I,R	RF	SM	T,F											
20	hadi		RF		T,F											

[1] Pre-processing: B= bias field correction, F= morphological filter to enhance small lesions, I= intensity normalization, R= resizing or resampling to a predefined grid, S= skull stripping, and T= transformation to a standard space.
[2] Post-processing: FP location based false positive reduction, G graph-based segmentation refinement, SL remove slices prone to false positives, and SM remove small segmentation results.
[3] DL indicates whether this method uses deep learning.
[4] Augmentation of training data: D= non-linear deforming, H= shearing, M= mirroring, R= rotating, S= scaling, T= translating/moving, and Y= generating synthetic lesions.
[5] Features used in the neural networks. Dim: 2D or 3D convolutions. Dil: dilated convolutions. BN: batch normalisation. Drop: dropout. MS: multi scale approaches (e.g. separate paths at different resolutions). LR: use of a learing rate schedule (e.g. reducing the learning rate during training). HN: hard negative mining. Ens: an ensemble of multiple networks.
* additional data from other sources was used to train this method.
† the convolutions are 2D, but the third dimension is processed within an RNN that incorporates all dimensions.

submitted methods employ some form of (deep) neural networks, including all methods in the top ten. Nevertheless, the use of deep learning methodology is not a guaranteed recipe for success, since a number of low-ranking methods use it as well.

Ensemble methods appear to do very well in this challenge. The methods of **sysu_media** (# 1), **nlp_logix** (# 3), and **k2** (# 5) use an ensemble of separately trained neural networks to achieve top-ranking results. Furthermore, the STAPLE algorithm that combines all methods or the top 4 ranking methods achieves good results as well. On the inter-scanner robustness ranking, both results of the STAPLE algorithm outperform all other participating methods. Combining the results of various methods has been performed in other challenges as well, for example in the brain tumor segmentation challenge (BRATS) [13]. However, in that challenge the combination of methods always outperformed all individual methods, whereas in this challenge the method of **sysu_media** remains the winner. This seems to be mainly caused by the good performance of **sysu_media** on the recall metric compared to the STAPLE results. Both STAPLE methods perform less well in recalling small lesions below the median size, as can be seen in Fig. 6.

The use of dropout during training is another characteristic of top-ranking methods. Random dropouts prevent units in neural networks from co-adapting too much [67] and introduces some redundancy in the network. A larger network trained with dropout might behave like an ensemble of smaller networks; and ensemble methods also rank at the top. However, the deep learning methods trained with dropout have a considerably lower inter-scanner rank (Table III). They drop more in the inter-scanner ranking than methods trained without dropout, suggesting that these methods might not generalize well to unseen data from unseen scanners; but instead only to unseen data from the same scanners as in the training data.

Selectively sampling WMH mimics, locations that resemble WMH but are not, or hard negative mining appears to be advantageous as well, since the three methods that apply it are amongst the top-ranking methods. When comparing the false positive maps of methods **nlp_logix** (Appendix C Figure 10), **nic-vicorob** (Appendix C Figure 11), and **misp** (Appendix C Figure 13) with that of the winner, **sysu_media** (Appendix C Figure 8); all three methods have less false positives (data not shown). However, this difference is not directly noticeable in any of the metrics in Table II, so the sampling strategy

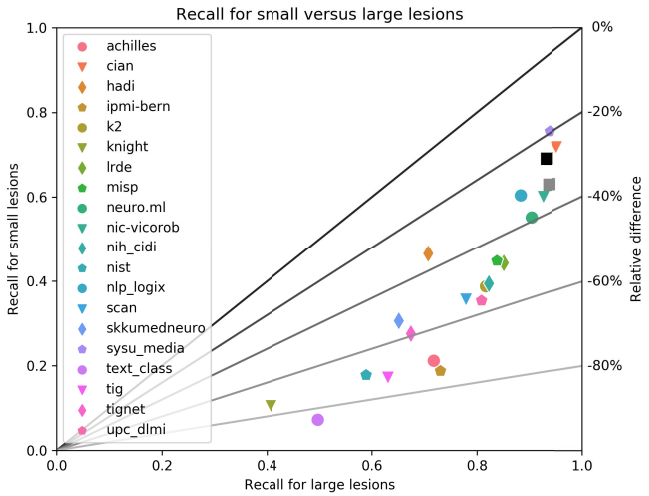


Fig. 6. Plot showing the recall of each method for small and large lesions. The right vertical axis indicates the relative difference for small lesions with respect to that of large lesions. Small lesions are defined as all lesions smaller than or equal to the median lesion volume per subject. Large lesions are all lesions larger than the median lesion volume per subject. The black and grey squares indicate the results of STAPLE applied on the top 4 or all methods, respectively.

might have had a minimal influence. A common location for false positive detections is the septum pellucidum, the area that separates both lateral ventricles. This can be seen in the third and fourth picture on the bottom row of Fig. 4. This area appears hyperintense on FLAIR, similar to WMH, but is never part of a WMH as can be seen in the top row of Fig. 4. Most top-ranking methods have no false positives in this area, whereas most lower-ranking methods do.

Implementing batch normalization, multi-scale processing, or using a learning rate schedule does not seem to influence the ranking of deep learning methods. The three methods that use dilated convolutions cluster together in the middle of the ranking, but whether that is attributable to the use of dilated convolutions or other factors is not sure.

Most deep learning methods that use 3D convolutions achieve a low ranking in the challenge. It could be that training 3D convolutional neural networks involved too many parameters, which could not be learned from the provided training data. Most FLAIR images were 2D multi-slice acquisitions (approximately $1 \times 1 \times 3$ mm voxels) with relatively few slices. Training 2D convolutional neural networks appears to work better in this case, but the methods of **cian**, **nic-vicorob**, and **misp** demonstrate that it was feasible to train 3D networks.

Regions with the highest false negative rates are located in regions with fewer WMH, as can be seen in the top and middle rows of Fig. 4. It appears that methods have issues finding WMH of which there are fewer training examples. This holds for all methods, as can be seen in the individual maps in Appendix C. Furthermore, the regions with high false negative rates usually have smaller WMH, for which the recall is lower compared to larger WMH (Fig. 6). It has been noted before that smaller WMH are harder to find and the proposed solution was to develop designated methods for small WMHs [68]. This has been adopted by the method of **nic-vicorob**, where a separate network reclassifies detected locations below a size of 30 voxels. Additionally, a selective sampling strategy might be used, combined with data augmentation, to provide more examples of small lesions during training. The method of **lrde** highlights small WMH as part of the pre-processing, but does not adapt the sampling strategy. Furthermore, method developers might need to make their methods less location-sensitive: not rejecting a WMH because it is at a location with low a priori probability. This might also be a strategy to reduce the number of false positive detections. These appear to coincide with the location of true positives, suggesting that methods more easily segment a false positive at locations with high a priori probability.

TABLE V

THE RE-COMPUTED RESULTS OF THE INTER-SCANNER ROBUSTNESS RANKING WHEN INCLUDING THE SIMULTANEOUS TRUTH AND PERFORMANCE LEVEL ESTIMATION (STAPLE) ALGORITHM APPLIED ON ALL METHODS OR ON THE TOP 4 RANKING METHODS. STAPLE OUTPERFORMS ALL METHODS AND THEREFORE THE RELATIVE RANKING VALUES CHANGE. HERE, STAPLE IS COMPARED TO THE TOP 3 METHODS IN THE ORIGINAL INTER-SCANNER RANKING IN TABLE III. THE SYMBOLS BETWEEN BRACKETS INDICATE WHETHER A TEAM IS RANKED ON THE SAME POSITION (−), LOWER (∨), OR HIGHER (∧) COMPARED TO THE ORIGINAL RANKING; WITH THE NEW POSITION INDICATED AS WELL

#	Team	Inter-scanner rank	
	STAPLE (top 4)		0.0152 (1)
	STAPLE (all)	0.0390 (1)	
1	ipmi-bern	0.0400 (∨ 2)	0.0402 (∨ 2)
2	sysu_media	0.0433 (∨ 3)	0.0434 (∨ 3)
3	achilles	0.0769 (∨ 4)	0.0768 (∨ 4)

The inter-scanner robustness ranking in the last column of Table III shows some remarkable changes in the ranking. The method of **ipmi-bern** becomes first, having the best inter-scanner robustness and putting **sysu_media** at the second place. Furthermore, the methods of **achilles**, **knight**, and **skkumedneuro** rank considerably higher. Despite the somewhat moderate performance on the individual metrics, these methods generalize well to unseen scanners and have robust performance; ranking very close to the winner. The top 10 of the inter-scanner ranking shows three non-deep learning methods, whereas none is present in the final ranking. The methods of **nic-vicorob** and **lrde** drop considerably in the inter-scanner ranking. Both methods perform less well on the images from the 3 T Philips Ingenuity (PET/MR) scanner that was not in the training data. Since only 10/110 test images originated from this scanner, it likely did not affect their overall ranking that much. The inter-scanner ranking of the **tig** and **tignet** methods shows a remarkable difference with the overall ranking. The **tignet** method, a neural network trained to replicate the results of the **tig** method, ranks close to the **tig** method in the overall ranking. In the inter-scanner ranking, the **tignet** method drops whereas the **tig** method rises.

No method performs best/worst on all individual metrics. Neither on the overall rankings nor on the inter-scanner rankings in Table III, ranking 0.0000 (overall best) nor 1.0000 (overall worst) are assigned to a method. Most room for improvement seems to be on the recall and F1 metrics.

Many methods fail to achieve a good score on these, which seems to be caused by methods missing small individual lesions. Missing one or a few small lesions does not contribute to a lower DSC, H95, nor lAVD, but does have a considerable influence on the recall and F1 metrics. Recent evidence shows that the presence and shape of small WMH can be of added value to further unravel the etiology and functional impact of WMH [69]. Furthermore, WMH location in strategic white matter tracts can explain cognitive dysfunctioning better than total WMH volume [4]. Hence, evaluating the recall and F1 metrics are of increasing importance for WMH segmentation methods.

Future developments in WMH segmentation might focus on improving the recall for small lesions and the inter-scanner robustness, especially on unseen data from unseen scanners. However, the current top ranking deep learning methods can already assist, or even replace, individual human observers in segmenting WMH.

After the results were presented at the MICCAI conference, a number of participants submitted an updated version of their method: **misp**, **neuro.ml**, **nih_cidi**, **sysu_media**, and **tig**. All methods showed an increased performance with respect to their original submission. Updated descriptions and results are available on the challenge website.

The WMH Segmentation Challenge remains open for new and updated future submissions.

ACKNOWLEDGMENT

The organizers thank T. Doeven for assisting with the manual segmentation of WMH. E. Puybareau and Y. Xu thank NVIDIA Corporation for donating a GeForce GTX 1080 Ti.

H. J. Kuijf and M. A. Viergever are with the Image Sciences Institute, UMC Utrecht, Utrecht University, Utrecht, The Netherlands (e-mail: h.kuijf@umcutrecht.nl).

J. M. Biesbroek, R. Heinen, and G. J. Biessels are with the Brain Center Rudolf Magnus, UMC Utrecht, Utrecht University, Utrecht, The Netherlands.

J. de Bresser is with the Department of Radiology, UMC Utrecht, Utrecht University, Utrecht, The Netherlands, and also with the Department of Radiology, LUMC, Leiden, The Netherlands.

S. Andermatt and S. Pezold are with the Department of Biomedical Engineering, University of Basel, Allschwil, Switzerland.

M. Bento is with the Departments of Radiology and Clinical Neuroscience, Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada.

M. Berseth is with NLP Logix, Jacksonville, FL USA.

M. Belyaev is with the Center for Neurobiology and Brain Restoration, Skolkovo Institute of Science and Technology, Moscow, Russia.

M. J. Cardoso is with the School of Biomedical Engineering and Imaging Sciences, King's College London, London, U.K., and also with the Centre for Medical Image Computing, University College London, London, U.K.

A. Casamitjana and V. Vilaplana are with the Signal Theory and Communications Department, Universitat Politècnica de Catalunya (BarcelonaTech), Barcelona, Spain.

D. L. Collins and M. Dadar are with the Montreal Neurological Institute, McGill University, Montreal, QC, Canada.

A. Georgiou is with the Computational Statistics and Machine Learning MSc Program, University College London, London, U.K.

M. Ghafoorian is with TomTom, Amsterdam, The Netherlands.

D. Jin and Z. Xu are with the Department of Radiology and Imaging Science, National Institutes of Health, Bethesda, MD USA.

A. Khademi is with the Image Analysis in Medicine Laboratory, Ryerson University, Toronto, ON, Canada.

J. Knight is with the University of Guelph, Guelph, ON, Canada.

H. Li is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, also with the School of Science and Engineering, University of Dundee, Dundee, U.K., and also with the Department of Computer Science, Technical University of Munich, Munich, Germany.

X. Lladó and S. Valverde are with the Research Institute of Computer Vision and Robotics, University of Girona, Girona, Spain.

M. Luna and S. H. Park are with the Department of Robotics Engineering, Daegu Gyeongbuk Institute of Science and Technology, Daegu, South Korea.

Q. Mahmood is with the Pakistan Institute of Nuclear Science and Technology, Islamabad, Pakistan.

R. McKinley and R. Wiest are with the Support Center for Advanced Neuroimaging, Institute for Diagnostic and Interventional Neuroradiology, Inselspital, University of Bern, Bern, Switzerland.

A. Mehrtash is with the Electrical and Computer Engineering Department, The University of British Columbia, Vancouver, BC, Canada, and also with the Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA USA.

S. Ourselin is with the School of Biomedical Engineering and Imaging Sciences, King's College London, London, U.K.

B.-Y. Park is with the Department of Electronic, Electrical and Computer Engineering, Sungkyunkwan University, Seoul, South Korea, and also with the Center for Neuroscience Imaging Research, Institute for Basic Science (IBS), Suwon, South Korea.

H. Park is with the Center for Neuroscience Imaging Research, Institute for Basic Science (IBS), Suwon, South Korea, and also with the School of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon, South Korea.

E. Puybareau is with the Research and Development Laboratory (LRDE), EPITA, Le Kremlin-Bicêtre Cedex, France.

L. Rittner is with the School of Electrical and Computer Engineering, University of Campinas, Campinas, Brazil.

C. H. Sudre is with the School of Biomedical Engineering and Imaging Sciences, King's College London, London, U.K., and also with the Centre for Medical Image Computing and the Dementia Research Centre, Institute of Neurology, University College London, London, U.K.

Y. Xu is with the Research and Development Laboratory (LRDE), EPITA, Le Kremlin-Bicêtre Cedex, France,

also with LTCI, Télécom ParisTech, Université Paris-Saclay, Saint-Aubin, France, and also with the Vision and Learning Representation Group, Huazhong University of Science and Technology, Wuhan, China.

G. Zeng and G. Zheng are with the Institute for Surgical Technology and Biomechanics, University of Bern, Bern, Switzerland.

J. Zhang is with the Department of Computing, University of Dundee, Dundee, U.K.

C. Chen is with the Memory Aging and Cognition Center, NUHS, Singapore.

W. van der Flier is with the Alzheimer Center, VU Amsterdam, Amsterdam, The Netherlands.

F. Barkhof is with the Department of Radiology and Nuclear Medicine, VU University Medical Center, Amsterdam, The Netherlands, and also with the UCL Institute of Neurology and Healthcare Engineering, University College London, London, U.K.

REFERENCES

[1] L. Pantoni, "Cerebral small vessel disease: From pathogenesis and clinical characteristics to therapeutic challenges," *Lancet Neurol.*, vol. 9, no. 7, pp. 689–701, 2010.

[2] N. D. Prins and P. Scheltens, "White matter hyperintensities, cognitive impairment and dementia: an update," *Nature Rev. Neurology*, vol. 11, no. 3, pp. 157–165, 2015.

[3] J. M. Wardlaw *et al.*, "Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration.," *Lancet. Neurol.*, vol. 12, no. 8, pp. 822–838, 2013.

[4] J. M. Biesbroek, N. A. Weaver, and G. J. Biessels, "Lesion location and cognitive impact of cerebral small vessel disease," *Clin. Sci.*, vol. 131, no. 8, pp. 715–728, 2017.

[5] M. E. Caligiuri, P. Perrotta, A. Augimeri, F. Rocca, A. Quattrone, and A. Cherubini, "Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: A review," *Neuro Inform.*, vol. 13, no. 3, pp. 261–276, 2015.

[6] H. Greenspan, B. V. Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, May 2016.

[7] T. Heimann *et al.*, "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1251–1265, Aug. 2009.

[8] K. Murphy *et al.*, "Evaluation of registration methods on thoracic CT: The EMPIRE10 challenge," *IEEE Trans. Med. Imag.*, vol. 30, no. 11, pp. 1901–1920, Nov. 2011.

[9] J. M. Wolterink *et al.*, "An evaluation of automatic coronary artery calcium scoring methods with cardiac CT using the orCaScore framework," *Med. Phys.*, vol. 43, no. 5, pp. 2361–2373, Apr. 2016.

[10] K. Sirinukunwattana *et al.*, "Gland segmentation in colon histology images: The glas challenge contest," *Med. Image Anal.*, vol. 35, pp. 489–502, 2017.

[11] M. Styner *et al.*, "3D segmentation in the clinic: A grand challenge II: MS lesion segmentation," *Midas J.*, pp. 1–6, Sep. 2008. [Online]. Available: http://hdl.handle.net/10380/1509

[12] O. Commowick *et al.*, "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 13650. [Online]. Available: http://www.nature.com/articles/s41598-018-31911-7

[13] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[14] A. M. Mendrik *et al.*, "MRBrains challenge: Online evaluation framework for brain image segmentation in 3T MRI scans," *Comput. Intell. Neurosci.*, vol. 2015, p. 1, Jan. 2015.

[15] J. M. F. Boomsma *et al.*, "Vascular cognitive impairment in a memory clinic population: Rationale and design of the 'utrecht-amsterdam clinical features and prognosis in vascular cognitive impairment' (TRACE-VCI) study," *JMIR Res. Protocols*, vol. 6, no. 4, p. e60, Apr. 2017.

[16] S. J. V. Veluw *et al.*, "Cortical microinfarcts on 3T MRI: Clinical correlates in memory-clinic patients," *Alzheimer's Dementia*, vol. 11, no. 12, pp. 1500–1509, 2015.

[17] J. Ashburner and K. J. Friston, "Voxel-based morphometry–The methods," *Neuro Imag.*, vol. 11, no. 6, pp. 805–821, Jun. 2000.

[18] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "elastix: A toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 196–205, Jan. 2010.

[19] D. Merkel, "Docker: Lightweight linux containers for consistent development and deployment," *Linux J.*, vol. 2014, no. 239, p. 5, 2014.

[20] W. Li, G. Wang, L. Fidon, S. Ourselin, M. J. Cardoso, and T. Vercauteren, "On the compactness, efficiency, and representation of 3D convolutional networks: Brain parcellation as a pretext task," in *Proc. Int. Conf. Inf. Process. Med. Imag. (IPMI)*, vol. 10265. Cham, Switzerland: Springer, 2017, pp. 348–360.

[21] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. (2017). "Rethinking atrous convolution for semantic image segmentation," [online]. Availabe: https://arxiv.org/abs/1706.05587

[22] A. Georgiou, (2017). *WMH Segmentation Challenge MICCAI 2017: Team Name-Achilles*. [Online]. Available: http://wmh.isi.uu.nl/results/achilles/

[23] S. Andermatt, S. Pezold, and P. Cattin, "Multi-dimensional gated recurrent units for the segmentation of biomedical 3D-data," in *Deep Learning and Data Labeling for Medical Applications*, G. Carneiro *et al.*, Eds. Cham, Switzerland: Springer, 2016, pp. 142–151.

[24] S. Andermatt, S. Pezold, and P. Cattin. (2017). *Multi-Dimensional Gated Recurrent Units for the Segmentation of White Matter Hyperintensites*. [Online]. Available: http://wmh.isi.uu.nl/results/cian/

[25] S. Andermatt, S. Pezold, and P. C. Cattin, "Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes, Eds. Cham, Switzerland: Springer, 2018, pp. 31–42.

[26] Q. Mahmood and A. Basit. (2017). *Automated Segmentation of White Matter Hyperintensities in Multi-Modal MRI Images Using Random Forests*. [Online]. Available: http://wmh.isi.uu.nl/results/hadi/

[27] G. Zeng and G. Zheng. (2017). *Deeply Supervised Multi-Scale Fully Convolutional Networks for Segmentation of White Matter Hyperintensities*. [Online]. Available: http://wmh.isi.uu.nl/results/ipmi-bern/

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI* (Lecture Notes in Computer Science), vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[29] A. Mehrtash and M. Ghafoorian. (2017). *Simurgh Team Method Description*. [Online]. Available: http://wmh.isi.uu.nl/results/k2/

[30] V. Fonov, A. C. Evans, K. N. Botteron, C. R. Almli, R. C. McKinstry, and D. L. Collins, "Unbiased average age-appropriate atlases for pediatric studies," *Neuro Imag.*, vol. 54, no. 1, pp. 27–313, 2011.

[31] J. Knight, G. Taylor, and A. Khademi. (2017). *Voxel-Wise Logistic Regression for White Matter Hyperintensity Segmentation in FLAIR MRI*. [Online]. Available: http://wmh.isi.uu.nl/results/knight/

[32] J. Knight, G. W. Taylor, A. Khademi, "Voxel-wise logistic regression and leave-one-source-out cross validation for white matter hyperintensity segmentation," *Magn. Reson. Imag.*, vol. 54, pp. 119–136, Dec. 2018.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2015, pp. 1–14.

[34] K.-K. Maninis, J. Pont-Tuset, P. A. Arbelàe, and L. V. Gool, "Deep retinal image understanding," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI* (Lecture Notes in Computer Science), S. Ourselin, L. Joskowicz, M. Sabuncu, G. Unal, and W. Wells, Eds. Cham, Switzerland: Springer, 2016.

[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[36] Y. Xu and T. Géraud author, É. Puybareau, I. Bloch, and J. Chazalon. (2017). *White Matter Hyperintensities Segmentation Using Fully Convolutional Network and Transfer Learning*. [Online]. Available: http://wmh.isi.uu.nl/results/lrde/

[37] Y. Xu, T. Géraud author, É. Puybareau, I. Bloch, and J. Chazalon, "White matter hyperintensities segmentation in a few seconds using fully convolutional network and transfer learning," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Lecture Notes in Computer Science), A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes, Eds. Springer, Cham, 2018, pp. 501–514.

[38] K. He, X. Zhang, S. Ren, and J. Sun. (Dec. 2015). "Deep residual learning for image recognition." [online]. Available: https://arxiv.org/abs/1512.03385

[39] M. Luna and S. H. Park. (2017). *3D Convolutional Neural Network With Skip Connections for WMH Segmentation*. [Online]. Available: http://wmh.isi.uu.nl/results/misp/

[40] K. Kamnitsas *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Imag. Anal.*, vol. 36, pp. 61–78, Feb. 2017.

[41] A. Safiullin. (2017). *NeuroML Team: Brief Description of the Solution*. [Online]. Available: http://wmh.isi.uu.nl/results/neuro-ml/

[42] S. Valverde *et al.*, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach," *Neuro Imag.*, vol. 155, pp. 159–168, Jul. 2017.

[43] S. Valverde *et al.* (2017). *White Matter Hyperintensity and Stroke Lesion Segmentation and Differentiation Using Convolutional Neural Networks*. [Online]. Available: http://wmh.isi.uu.nl/results/nic-vicorob/

[44] D. Jin. (2017). *WMH Segmentation Method Description—NIH_CIDI*. [Online]. Available: http://wmh.isi.uu.nl/results/nih_cidi/

[45] M. Dadar *et al.*, "Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging," *Neuro Imag.*, vol. 157, pp. 233–249, 2017.

[46] M. Dadar *et al.*, "Validation of a regression technique for segmentation of white matter hyperintensities in alzheimer's disease," *IEEE Trans. Med. Imag.*, vol. 36, no. 8, pp. 1758–1768, Aug. 2017.

[47] M. Dadar, V. S. Fonov, and D. L. Collins. (2017). *Automatic Multi-Modality Segmentation of White Matter Hyperintensities Using a Random Forests Classifier*. [Online]. Available: http://wmh.isi.uu.nl/results/nist/

[48] M. Ghafoorian *et al.*, "Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities," *Sci. Rep.*, vol. 7, no. 1, p. 5110, 2017.

[49] M. Berseth. (2017). *WMH Segmentation Challenge, MICCAI 2017*. [Online]. Available: http://wmh.isi.uu.nl/results/nlp_logix/

[50] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. 6th Int. Conf. Learn. Represent.*, 2016.

[51] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[52] R. McKinley, A. Jungo, R. Wiest, and M. Reyes, (2017). *Pooling-Free Fully Convolutional Networks With Dense Skip Connections for Semantic Segmentation, With Application to Segmentation of White Matter Lesions*. [Online]. Available: http://wmh.isi.uu.nl/results/scan/

[53] B.-Y. Park, M. J. Lee, and H. Park. (2017). *WMH Segmentation Challenge at MICCAI 2017: Brief Description of the Method*. [Online]. Available: http://wmh.isi.uu.nl/results/skkumedneuro/

[54] H. Li, G. Jiang, L. Zhao, R. Wang, J. Zhang, and W.-S. Zheng. (2017). *Automatic White Matter Hyperintensity Segmentation Via Two-Channel U-Net*. [Online]. Available: http://wmh.isi.uu.nl/results/sysu_media/

[55] H. Li *et al.*, "Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images," *Neuro Imag.*, vol. 183, pp. 650–665, Dec. 2018.

[56] M. Bento, R. de Souza, R. Lotufo, R. Fraynea, and L. Rittner. (2017). *WMH Segmentation Challenge: A Texture-Based Classification Approach (ID: Textclass)*. [Online]. Available: http://wmh.isi.uu.nl/results/text_class/

[57] M. Bento, R. de Souza, R. Lotufo, R. Frayne, and L. Rittner, "WMH segmentation challenge: A texture-based classification approach," in *Brainlesion: Glioma, Multiple Sclerosis*, A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes, Eds. Cham, Switzerland: Springer, 2018, pp. 489–500.

[58] C. H. Sudre, M. J. Cardoso, W. H. Bouvy, G. J. Biessels, J. Barnes, and S. Ourselin, "Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 2079–2102, Oct. 2015.

[59] C. Sudre. (2017). *Team TIG—WMH Challenge*. [Online]. Available: http://wmh.isi.uu.nl/results/tig/

[60] (2017). *TIGNet—WMH Challenge*. [Online]. Available: http://wmh.isi.uu.nl/results/tignet/

[61] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for, Volumetric Medical Image Segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[62] A. Casamitjana, M. Combalia, and I. Sánchez, and V. Vilaplana. (2017). *Augmented V-Net for White Matter Hyperintensities Segmentation*. [Online]. Available: http://wmh.isi.uu.nl/results/upc_dlmi/

[63] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.

[64] B. L. Welch, "The generalization ofstudent's' problem when several different population variances are involved," *Biometrika*, vol. 34, nos. 1–2, pp. 28–35, 1947.

[65] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. (Jun. 2018). "Do CIFAR-10 classifiers generalize to CIFAR." [Online]. Available: http://arxiv.org/abs/1806.00451

[66] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Imag. Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[67] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[68] M. Ghafoorian *et al.*, "Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease," *Med. Phys.*, vol. 43, no. 12, pp. 6246–6258, Dec. 2016.

[69] J. d. Bresser, H. J. Kuijf, K. Zaanen, M. A. Viergever, J. Hendrikse, and G. J. Biessels, "White matter hyperintensity shape and location feature analysis on brain MRI; Proof of principle study in patients with diabetes," *Sci. Rep.*, vol. 8, no. 1, p. 1893, Dec. 2018.